

DSDBASE: a consortium of native and modelled disulphide bonds in proteins

A. Vinayagam¹, G. Pugalenth¹, R. Rajesh and R. Sowdhamini*

National Centre for Biological Sciences, UAS-GKVK Campus, Bangalore 560065, India

Received July 22, 2003; Revised August 16, 2003; Accepted September 3, 2003

ABSTRACT

DSDBASE is a database of disulphide bonds in proteins, which provides information on native disulphides and those that are stereochemically possible between pairs of residues for all known protein structural entries. The modelling of disulphides has been performed, using MODIP, by the identification of residue pairs that can strainlessly accommodate a covalent cross-link. We also assess the stereochemical quality of the covalent cross-link and grade them appropriately. One of the potential uses of the database is to design site-directed mutants in order to enhance the thermal stability of a protein. The proposed sites of mutations can be viewed specifically with respect to active sites of enzymes and across physiological dimers. The occurrence of native and modelled disulphides increases the dimensions of the database enormously. This database can also be employed for proposing three-dimensional models of disulphide-rich short polypeptides. The database can be accessed from <http://www.ncbs.res.in/~faculty/mini/dsdbase/dsdbase.html>. Supplementary information can be accessed from <http://www.ncbs.res.in/~faculty/mini/dsdbase/nar/suppl.htm>.

INTRODUCTION

Disulphide bonds are Cys–Cys covalent linkages that connect different parts of a protein. Disulphides have been recorded in 29% of protein structures (5737 of 19 612 entries corresponding to the April 2003 release) in the protein data bank (PDB) (1,2). A large proportion of peptides in the sequence databases are rich in disulphides with no structural information available and several of them are bioactive. Structural information of such important molecules can be extrapolated from homology to protein structural entries that are considered in a database of disulphides. The size of the disulphide database can be enhanced substantially by including substructures where SS bonds can be modelled amongst pairs of residues in a protein. MODIP (3) is the procedure employed to include putative

disulphide cross-links where disulphides are modelled using stereochemical criteria (see DSDBASE at <http://www3.oup.co.uk/nar/database/a/>). In this paper, we report the availability of a database that includes native and modelled disulphide cross-links for all known entries in the protein structural data bank.

Enhancement of protein thermal stability is expected due to the introduction of new disulphide bonds by site-directed mutagenesis (4). In order to design point mutations, the availability of information on possible sites for the strainless introduction of disulphides for a large number of proteins is highly desirable. This will form another useful application of the database.

GENERAL FEATURES OF THE DATABASE

DSDBASE comprises the positions and cross-link stereochemistry of modelled and native disulphide bonds that connect protein substructures. In the current modelling approach, all possible residue pairs are examined for their stereochemical compatibility to accommodate disulphides. Cα–Cα and Cβ–Cβ distance criteria are employed to screen appropriate residue pairs. Disulphide bonds are modelled by geometric fixing of sulphur atoms, once the distance compatibility is achieved [please see (3) for details]. Modelled disulphide bonds are graded according to their stereochemical parameters that describe the geometry of bridges such as side chain torsion angles. Native disulphide bonds are marked to differentiate them from the modelled ones. The loop size or spatial distance between two residues participating in a disulphide bond, which may facilitate choosing the best possible position for the introduction of disulphide bonds, is also recorded. Separate datasets are available for A, B, C grades and native disulphides (please see below).

ACCESS TO THE DATABASE

DSDBASE can be accessed from <http://www.ncbs.res.in/~faculty/mini/dsdbase/dsdbase.html>. The database considers all PDB entries in order to increase the chances of access to information on disulphide bond formation. The inclusion of all PDB entries (PDB April 2003 release), and sometimes all models within an NMR entry, seems valuable since we find

*To whom correspondence should be addressed. Tel: +91 80 3636421; Fax: +91 80 3636462; Email: mini@ncbs.res.in
Present address:

A. Vinayagam, Department of Molecular Biophysics, German Cancer Research Centre (DKFZ), Im Neuenheimer Feld 280, D-69120 Heidelberg, Germany

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

that closely related proteins contribute to DSDBASE in a manner that provides complementary information. A simple keyword search and PDB code search options are available in order to reach the specific protein. All possible pairs of residues that can accommodate a disulphide bond are listed along with the stereochemistry of the cross-link. A search program is provided online for probing the database for particular disulphide bond connectivity and all substructural motifs that can accommodate the restraints can be recognized. Options for relaxing loop size and inter-disulphide positions are also available.

FEATURES AND INTERFACED TOOLS

Sites of mutations

All the 'mutant' PDB files can be visualized over the Web using RASMOL (5) and CHIME (MDL Information Systems, Inc.) graphic interfaces. Sites that are both Cys residues could correspond to native disulphides; of these, a subset are perhaps annotated in the PDB file; some native disulphides might contain inherent strain due to functional requirements if they are in the active site of thiol oxidoreductases; Cys-Cys pairs of the above types are distinguished in the list of sites (please see supplementary information for a sample output). MODIP (3) is available online and can be applied to new protein structures that are not yet recorded in PDB or for particular multimeric assemblies for the examination of sites where disulphides can be introduced strainlessly.

Functional and structural information of protein substructures

The intracellular environment is heavily reducing and therefore intracellular proteins are less likely to retain disulphide bonds despite overall stereochemical suitability to accommodate such cross-links. Cellular localisation is predicted using SUBLOC (6) and this information is provided to consider the feasibility of disulphide bonds present in such proteins. In addition, the following features have been considered.

(i) From sources such as the enzyme data bank (7), consolidated information is provided for enzymes about active site residues in addition to those provided in the PROCAT database (8) and PDBSUM (9). The active-site residues and sites near the active site (within 5 Å distance) are highlighted (please see Supplementary Material for a sample output). This can also be visualized through Rasmol/Chime links. Enzyme entries can also be searched by their EC numbers.

(ii) For NMR-determined protein structures there are options to search in a specific model or all the models in the ensemble of conformations reported in the PDB entry. Information about the clustering of models within such PDB entries is provided and it is possible to select a representative structure (10).

(iii) Given the coordinates of individual protomers of protein dimers, the online version of MODIP can suggest inter-protomer disulphide cross-links that are relevant to proteins that occur as physiological dimers and can increase thermal stability and activity.

(iv) The sulphur coordinates and stereochemical parameters for all the possible sulphur positions are also provided.

Table 1. Number of substructures recorded in DSDBASE

	Full database	Non-redundant (nr) database	Native of nr database
No. of proteins	19 612	2849	766
No. of SS bonds	2 385 617	149 892	2170
Grade ^a A	307 003 (12%)	19591 (13%)	1321 (61%)
Grade B	633 431 (26%)	39 804 (27%)	341 (16%)
Grade C	728 374 (30%)	90 497 (60%)	508 (23%)
Grade D ^b	716 809 (29%)	–	–

^aPlease see text and Sowdhamini *et al.* (3) for explanation of grade.

^bC α -C α and C β -C β distances were compatible but sulphur could not be fixed geometrically.

Modelling disulphide-rich polypeptides

It is possible to search the entire database using particular disulphide bond connectivity involving one or more disulphide bonds. The search procedure examines the individual polypeptide chains in the database for compatibility with the loop size and the spacing between inter-disulphide positions. A user-defined relaxation is permitted in the dimension of loop size as well as the inter-disulphide spacing. Compatible substructures from different proteins that satisfy the query disulphide bond connectivity are projected as output.

(i) There is an option to examine the primary structural compatibility between the query peptide and the substructural hits internally through MALIGN (11). Substructures are ranked in order of decreasing sequence similarity with the query polypeptide (please see supplementary information for a sample output).

(ii) A separate option is provided on the search result page to filter the hits on the basis of user-defined sequence identity cut-off. This helps to reduce the extent of redundancy in the hits.

Other features such as the crystallographic resolution for PDB entries, predicted sub-cellular location of the protein whose substructure is included in the database, the positions of native disulphide bonds and the positions of redox-active SS bonds, can be considered in choosing substructures for modelling.

DATABASE STATISTICS

19 612 protein structural entries have been examined for the position of native and modelled disulphides. The inclusion of modelled disulphides leads to an increase in the size of the database of 98%. DSDBASE records 2 385 617 protein substructures that have stereochemical compatibility to accommodate disulphide bonds (Table 1). Usually for a protein of 200 residues, ~45 residue pairs were stereochemically compatible to accommodate modelled disulphide bonds.

A vast majority of the annotated native disulphides (75%) could be modelled very well: 19 106 of 25 296 annotated SS bonds were 'modelled' as Grade A disulphides. Some other native SS bonds could be inherently strained due to functional constraints, like those involved in thiol oxidoreductase activity. 47 518 disulphides were identified from the non-redundant set (25% sequence identity cut-off) of 2849 protein chains (12) and are recorded in DSDBASE.

OUTLOOK

DSDBASE would be periodically updated with new entries in the protein data bank (1,2). MODIP has been widely used for the rational design of site-directed mutagenesis experiments leading to enhanced protein thermal stability (13–19). The recent version of MODIP (20) examines short contacts that might result due to the inclusion of disulphides in a protein fold. By making the MODIP program available online, any protein structure can be queried for possible sites to accommodate disulphides. We have provided graphical interfaces for viewing modelled disulphide bonds. In addition, we have incorporated biochemical data such as the predicted cellular localisation and the presence of redox-active disulphides. We have also considered spatial positions of suggested sites with respect to the enzyme active site, NMR ensembles and the occurrence of disulphides across physiological dimers.

We have recently benchmarked the search procedure for querying the database using SS bond connectivity from peptides of known structure (R. Rajesh, A. Vinayagam, G. Pugalenth and R. Sowdhamini, manuscript submitted). This approach gives rise to models close to the experimental structure in ~60% of the cases without any other experimental information other than amino acid sequence and disulphide bond connectivity. A curated database of disulphide cross-links of all known protein structural entries is a rich resource for further predictions and should be of value to biochemists and biologists.

ACKNOWLEDGEMENTS

We thank Professor Balaram for initiating this idea. This research is supported by the award of International Senior Fellowship in Biomedical Sciences to R.S. from the Wellcome Trust, UK. A.V. was supported by the Wellcome Trust. G.P. and R.R. are currently supported by the Wellcome Trust. We also thank NCBS for infrastructural support.

REFERENCES

- Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, E.F., Jr, Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.*, **25**, 535–542.
- Berman, H.M., Battistuz, T., Bhat, T.N., Bluhm, W.F., Bourne, P.E., Burkhardt, K., Feng, Z., Gilliland, G.L., Iype, L., Jain, S. *et al.* (2002) The Protein Data Bank. *Acta Crystallogr. D*, **58**, 899–907.
- Sowdhamini, R., Srinivasan, N., Shoichet, B., Santi, D.V., Ramakrishnan, C. and Balaram, P. (1989) Stereochemical modelling of disulfide bridges: Criteria for introduction into proteins by site-directed mutagenesis. *Protein Eng.*, **3**, 95–103.
- Wetzel, R., Perry, L.J., Baase, W.A. and Becktel, W.J. (1988) Disulfide bonds and thermal stability in T4 lysozyme. *Proc. Natl Acad. Sci. USA*, **85**, 401–405.
- Sayle, R.A. and Milner-White, E.J. (1995) RASMOL: biomolecular graphics for all. *Trends Biochem. Sci.*, **20**, 374–376.
- Hua, S. and Sun, Z. (2001) Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, **17**, 721–728.
- Bairoch, A. (1993) The ENZYME data bank. *Nucleic Acids Res.*, **21**, 3155–3156.
- Wallace, A.C., Borkakoti, N. and Thornton, J.M. (1997) TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases: application to enzyme active sites. *Protein Sci.*, **6**, 2308–2323.
- Laskowski, R.A. (2001) PDBsum: summaries and analyses of PDB structures. *Nucleic Acids Res.*, **29**, 221–222.
- Kelley, L.A., Gardner, S.P. and Sutcliffe, M.J. (1996) An automated approach for clustering an ensemble of NMR-derived protein structures into conformationally-related subfamilies. *Protein Eng.*, **9**, 1063–1065.
- Johnson, M.S., Overington, J.P. and Blundell, T.L. (1993) A structural basis for sequence comparisons; an evolution of scoring methodologies. *J. Mol. Biol.*, **233**, 735–752.
- Hobohm, U., Scharf, M., Schneider, R. and Sander, C. (1992) Selection of representative protein data sets. *Protein Sci.*, **1**, 409–417.
- Gokhale, R.S., Agarwalla, S., Francis, V.S., Santi, D.V. and Balaram, P. (1994) Thermal stabilization of thymidylate synthase by engineering two disulfide bridges across the dimer interface. *J. Mol. Biol.*, **235**, 89–94.
- Farzan, M., Choe, H., Desjardins, E., Sun, Y., Kuhn, J., Cao, J., Archambault, D., Kolchinsky, P., Koch, M., Wyatt, R. and Sodroski, J. (1998) Stabilization of human immunodeficiency virus type 1 envelope glycoprotein trimers by disulfide bonds introduced into the gp41 glycoprotein ectodomain. *J. Virol.*, **72**, 7620–7625.
- Topham, C.M., Mouldous, L., Poda, G., Maigret, B. and Meunier, J.C. (1998) Molecular modelling of the ORL1 receptor and its complex with nociceptin. *Protein Eng.*, **11**, 1163–1179.
- Velanker, S.S., Gokhale, R.S., Ray, S.S., Gopal, B., Parthasarathy, S., Santi, D.V., Balaram, P. and Murthy, M.R. (1999) Disulfide engineering at the dimer interface of *Lactobacillus casei* thymidylate synthase: crystal structure of the T155C/E188C/C244T mutant. *Protein Sci.*, **8**, 930–933.
- Gale, A.J., Xu, X., Pellequer, J.-L., Getzoff, E.D. and Griffin, J.H. (2002) Interdomain engineered disulfide bond permitting elucidation of mechanisms of inactivation of coagulation factor Va by activated protein C. *Protein Sci.*, **11**, 2091–2101.
- Ivens, A., Mayans, O., Szadkowski, H., Jurgens, C., Wilmanns, M. and Kirschner, K. (2002) Stabilization of a (β/α)8-barrel protein by an engineered disulfide bridge. *Eur. J. Biochem.*, **269**, 1145–1153.
- Pikkemaat, M.G., Linssen, A.B.M., Berendsen, H.J.C. and Janssen, D.B. (2002) Molecular dynamics simulations as a tool for improving protein stability. *Protein Eng.*, **15**, 185–192.
- Dani, V.S., Ramakrishnan, C. and Varadarajan, R. (2003) MODIP revisited: re-evaluation and refinement of an automated procedure for modeling of disulphide bonds in proteins. *Protein Eng.*, **16**, 187–193.